International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 3 Issue 2, February - 2014

A Review on Techniques for Generation of Knowledge in Scientific Database

Ms. Mrunali L. Vaidya
M. Tech. Department of Computer Science & Engg.
Priyadarshini Bhagwati Chaturvedi College of Engg.
Nagpur, India

Prof. M. S. Chaudhari
Department of Computer Science & Engg.
Priyadarshini Bhagwati Chaturvedi College of Engg.
Nagpur, India

***Abstract*— At present, the stored information is increasing tremendously day by day. The sudden increase in the amount of texts on the web, it was almost impossible for people to keep up-to-date information. Knowledge generation from textual database referred generally to the process of extracting interesting or non-retrieval patterns or knowledge from unstructured text documents. Using the technique such as information extraction, information retrieval, natural language processing, text mining and social network analysis can be easily found from the corpus of documents set. Knowledge generation in databases is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data. The development of proposed work is the acquirization or selection of target data set, integration and checking of data set, data cleaning, preprocessing and development of transformation model and selection of algorithm which gives generated knowledge as result interpretation, visualization, testing, verification and maintenance.**

Keywords—Knowledge generation, Corpus, Text mining, Social Network Analysis, Information Eextraction, Information Retrieval.

I. INTRODUCTION

The aim of data mining is the extraction of implicit knowledge from huge databases, by investigating patterns and regularities in data. Graph Mining is the subarea which focuses on mining from graph datasets. The contribution of data mining researchers to social network analysis focuses on two aspects. The first one is the extraction of models by analyzing examples of networks, in which data mining is used to identify some critical parameters, based only on structural and not semantically properties, that describe the microscopic evolution of some social networks[2]. The second aspect is focused on analyzing the temporal evolution of the network.

Data Mining and Knowledge generation in Databases are terms used interchangeably. There are some terms often used are data or information harvesting, data archaeology, functional dependency analysis, knowledge extraction and data pattern analysis. A high level definition of Data Mining is the non-trivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data. Data mining is not only a simple process but also it has the problem that there is no tool that can do job in automatic way. Data mining [19] can be aided by tools, but it requires both human data mining expertise and human domain expertise. Hence, large number of operations are present in data mining process each of which are supported by various technologies, like rule induction, neural networks, conceptual clustering. In real world applications, information extraction requires the co-operative use of several data mining operations and methods. Data mining and generation of knowledge in databases play vital role to interact specially focused on scientific database which are essential for analysis exploration operation. Data mining techniques is needed for computer-driven data exploration which used for future work and also for different database in new way. Query Specification at abstract level which also facilitates exploration of data for problems which arise because high dimensionality would otherwise prove to be very difficult to explore by humans, inspite this difficulty of use for efficiency issues with SQL. The query formulation problem is defined that has not accepted much more attention in database system.

Data mining is typically not used as a business system delivery technology. Rather it is an extremely powerful and effective set of technologies for analyzing and clustering data which can be used to form [9] the basis of a system. Knowledge Discovery and Data Mining (KDD) is an interdisciplinary area which focuses upon certain methodologies for extracting useful knowledge from data. The ongoing rapid growth of online data due to the Internet access and the widespread use of databases have created an immense and large need for KDD methodologies. The[16] challenge of extracting knowledge from data draws upon research in statistics, databases, pattern recognition, machine learning, data visualization, optimization, and high-performance computing, to deliver advanced business intelligence and web discovery solutions.

Web mining is the application of data mining techniques which is used to discover patterns from the Web. According to analysis targets, web mining can be divided into three different types, which are Web usage mining, Web content mining and Web structure mining Web usage mining is the process of extracting useful information from server logs e.g. use Web usage mining is the process of finding out what users are looking for on the Internet. Some users solely depend on the textual data, whereas some others might be interested in other multimedia data. Another application of data mining is web usage mining to generate usage patterns which are interesting from web data, so as to understand the web based application needs. The identity or origin of internet users is captured by usage data along with the browsing behavior of user at certain websites. Web content mining is the mining, extraction and integration of useful data, [4] information and

knowledge from Web page content. The heterogeneity and the lack of structure that allows much of the ever-expanding and immense increasing information sources on the World Wide Web, such as hypertext documents, makes automated discovery, organization, and search and indexing tools of the Internet and the World Wide Web.

II. OVERVIEW ON KNOWLEDGE GENERATION IN DATABASES

Knowledge generation in Databases brings together current research on the exciting problem of generating useful and interesting knowledge in databases. Scientific database are corpus file collection which file having some scientific text or collection of word from which we have to generate the knowledge and find the best suitable file from repository. Knowledge generation in Databases is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data. It spans many different approaches to discovery, including inductive learning, Bayesian statistics, and semantic query optimization, knowledge acquisition for expert systems, information theory, and fuzzy sets.

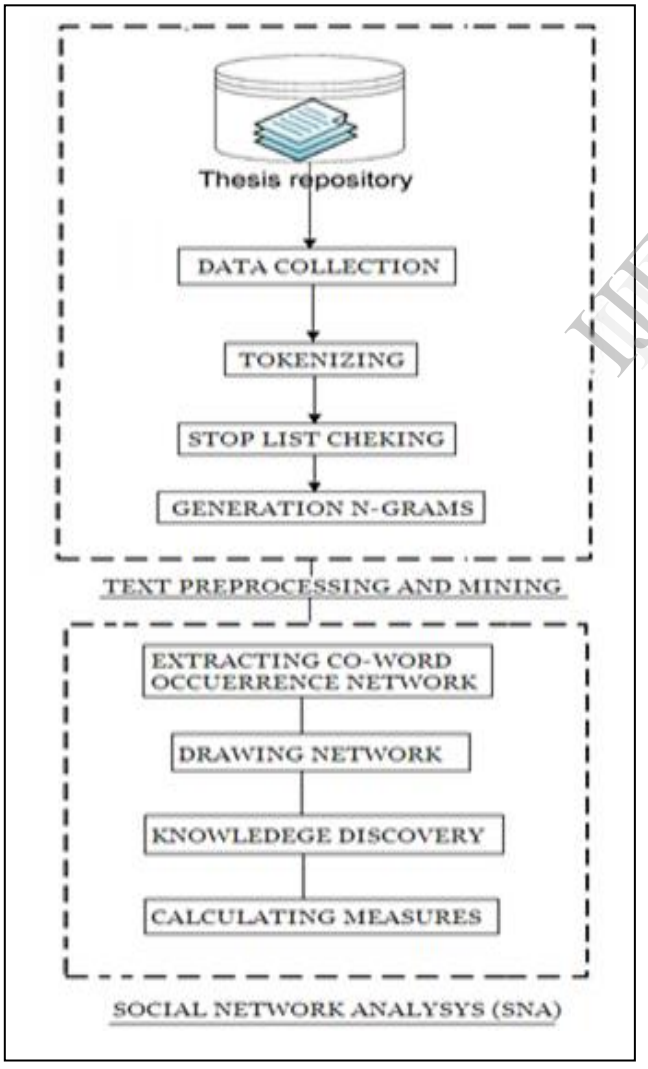

Fig. 1. Basic block diagram for generation of knowledge

The rapid growth in the number and size of databases creates a need for tools and techniques for intelligent data understanding. Relationships and patterns in data may enable a manufacturer to discover the cause of a persistent disk failure or the reason for consumer complaints. But today's databases hide their secrets beneath a cover of overwhelming detail. The task of uncovering these secrets is called "discovery in databases." This loosely defined subfield of machine learning is concerned with discovery from large amounts of possible uncertain data. Its techniques range from statistics to the use of domain knowledge to control search.

The knowledge generation goals are defined by (1) Verification (2) Generation. Verification states that system is limited to verify hypothesis of user. With generation of some patterns, the system autonomously finds new patterns. Further it gets divided into prediction, so that system can finds different patterns for predicting the entities and description, so that system can able to finds patterns which are used for user presentation which is in the form that human can understand. To generate the knowledge, keywords will be extracted from text records which characterize as a chain of one or more words. The continuous text should be tokenized to words, phrases, symbols and other elements called tokens and the extracted words are checked with the stop list to exclude unnecessary words. The N-grams are checked and meaningful ones are replaced with singular tokens in text. The text will be represented as a network and co-word occurrence will be extracted to calculate adjacency matrix of network and finally generate the knowledge.

III. RESEARCH CHALLENGES

- For large databases, develop mining algorithm for classification, clustering, analysis, modification and deviation detection. There are a much differences between performance and accuracy as one dependent to the fact that data solely resides primarily on disk or on a server and cannot fit in main memory.
- The mining algorithm can operate on a database for building of some specific techniques for encoding metadata and can be easily understood. Thus the more information can be asked by knowledge generation techniques effectively from user.
- While operating in a very large datasets size environment is a blessing against over fitting problems, data mining systems need to guard against fitting models to. As a huge search space is explored by a program for different data sets, this problem become more significant.
- Modeling of new data mining algorithm are able to account for structure as well as extracting more complex related between fields.
- Development of data mining operations that account for knowledge of data and exploitation such knowledge in reduction search that can account for costs and profits, which are robust against unassurity and missing data problems. Bayesian methods and decision analysis provide the basic foundational framework [19].

IV. RELATED WORK

The objective behind this paper is to overcome the problem of variable terminology by the aid of concept-based information retrieval.[1] The work is done on systematic generation of concept maps with the help of text mining techniques. They have used GetSmart software package to draw conceptual maps, and develops a publicly accessible repository of concept maps to enable sharing of the knowledge.

Decision tree is used to find bigrams that occur nearby. They have evaluated his approach using the sense-tagged corpora from the 1998 SENSEVAL word sense [2] disambiguation exercise. They showed that bigrams are powerful features for performing word sense disambiguation. They have also proved that an effortless decision tree where each node checks whether or not a particular bigram occurs near the ambiguous word results in accuracy comparable with state-of-the-art methods.

A novel methodology to extract core concepts from text corpus [3], methodology is based on text mining and social network analysis. At the text mining phase the keywords are extracted by tokenizing, removing stop-lists and generating N-grams. Network analysis phase includes co-word occurrence extraction, network representation of linked terms and calculating centrality measure.

Business databases that have accumulated over many [9] years of business records typically contain a wealth of hidden knowledge that could potentially be utilized by management to make better informed business decisions. They described a framework for knowledge discovery from business databases and demonstrate how the discovered knowledge can be put into good use. The proposed framework uses an application that integrates data mining processes with online analytical processing (OLAP).

There is another proposed algorithm to extract knowledge using predictive Apriori [10] and distributed grid based Apriori algorithms for association rule mining. They presented the implementation of an association rules discovery data mining task using Grid technologies. As a result of implementation with a comparison of classic Apriori and distributed Apriori.

Information Extraction (IE) and knowledge discovery in [6] databases (KDD) were both useful approaches for discovering information in textual corpora, but they have some deficiencies. The aim is to provide a new high-quality information extraction methodology and, at the same time, to improve the performance of the underlying extraction system.

Text mining algorithms suggested the overwhelming increase in the amount [4] of texts on the web, it was almost impossible for people to keep abreast of up-to-date information. Text mining algorithms was used to guarantee the quality of extracted knowledge. However, the extracted patterns using text or data mining algorithms or methods lead to noisy patterns and inconsistency.

The general framework of text mining consists of two different components, [5] text refining that transforms free form text documents into an intermediate form and knowledge distillation that deduces patterns or knowledge from intermediate form. Intermediate form (IF) can be semi structured such as the conceptual graph representation or structured such as the relational data representation. IF can be document based, here each entity represents the document or concept based, here each entity represents an object or concepts of interests in specific domain. The different techniques are included such as summarizing text, categorization, Naive Bayesian Classifier, Nearest Neighbour Classifier, Clustering etc.

TABLE I. A SURVEY ON DIFFERENT TECHNIQUES

Author	Title /Year of Publication	Techniques/Rules	Description
Ammar Jalalimanesh	Knowledge Discovery in Scientific Databases UsingText Mining and Social Network Analysis, 2012.	GetSmart software package	To draw concept maps, and established a publicly accessible repository of concept maps to enable sharing of the knowledge
M.Sukanya ,S.Biruntha	Techniques on Text Mining,2012	Naive Bayesian classifier, clustering, information extraction	To analyse unstructured text and to group similar documents, probabilistic relationship between different categories
Vaishali Bhujade ,N.J.Janwe	Knowledge discovery in text mining techniques using association rule extraction,2011	EART extraction association rule	To select most important keywords or features to form association rule
Christina Feilmayr	Text Mining-Supported Information Extraction,2011	TEMIE integrating data mining into information extraction	Based on various statistical and machine learning methods to meet requirement of different IE phases
R. Sumithra, Dr (Mrs). Sujni Paul	Using distributed apriori association rule and classical apriori mining algorithms for grid based knowledge discovery,2010	Apriori algorithm	To improve the efficiency of the level wise generation of frequent itemsets

V. CONCLUSION

After the investigation of all the techniques for generation of knowledge, it can be concluded that though various techniques such as web mining, data mining, information extraction, information retrieval etc. used for generation of knowledge, but still they are not efficient to generate unknown information of large databases. So for this more efficient approach is to use combined techniques such as text mining and social network analysis to generate the unknown information.

REFERENCES

[1] Ammar Jalalimanesh, "Knowledge Discovery in Scientific Databases UsingText Mining and Social Network Analysis",IEEE Conf. on Control, Systems and Industrial Informatics (ICCSII),PP.46-49,September 23-26,2012.

[2] Zhen Guo, Zhongfei (Mark) Zhang et.al, "A Two-Level Topic Model towards Knowledge Discovery from Citation Networks", IEEE Transaction on knowledge and data engineering, PP.1-30, 2013.

[3] K. M. Sam, C. R. Chatwin, "Ontology-Based Text-Mining Model For Social Network Analysis",PP.226-231,IEEE 2012.

[4] Mubarak Albathan, Yuefeng Li et.al, "Using Patterns Co-occurrence Matrix for Cleaning Closed Sequential Patterns for Text Mining", IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology,PP.201-205,2012.

[5] M.Sukanya,S.Biruntha, "Techniques on Text Mining",IEEE International Conf. on Advanced Communication Control and Computing Technologies (ICACCCT), PP.269-271,2012.

[6] Christina Feilmayr, "Text Mining-Supported Information Extraction", 22nd International Workshop on Database and Expert Systems Applications, PP.217-221,2011.

[7] Xin Guo, Yang Xiang, Qian Chen, "A Vector Space Model Approach to Social Relation Extraction from Text Corpus", Eighth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)PP.1756-1759,2011.

[8] Vaishali Bhujade ,N.J.Janwe, "Knowledge discovery in text mining techniques using association rule extraction",international conf.on computational intelligence and communication system,PP.498-502, 2011.

[9] A.C.M. Fong, "A Generalized Framework for Knowledge Discovery in Business Environments", Second International Conference on Communication Systems, Networks and Applications, PP.48-51,2010.

[10] Mrs. R. Sumithra, Dr (Mrs). Sujni Paul, "Using distributed apriori association rule and classical apriori mining algorithms for grid based knowledge discovery", Second International conference on Computing, Communication and Networking Technologies, PP.1-5, 2010.

[11] M. Fritsch and M. Kauffeld-Monz, "The impact of network structure on knowledge transfer: an application of social network analysis in the context of regional innovation networks," The Annals of Regional Science,vol. 44, PP. 21-38, 2010.

[12] T. Opsahl, et al., "Node centrality in weighted networks: Generalizing degree and shortest paths," Social Networks, vol. 32, PP. 245-251, 2010.

[13] N. Santoro, et al., "Time-Varying Graphs and Social Network Analysis: Temporal Indicators and Metrics," 2011.

[14] J. Jayabharathy , S. Kanmani , "Document Clustering and Topic Discovery based on Semantic Similarity in Scientific Literature" ,IEEE,PP.425-429,2011.

[15] Michele Coscia, Fosca Giannotti, Ruggero Pensa, "Social Network Analysis as Knowledge Discovery process", IEEE,Advances in Social Network Analysis and Mining, PP. 279-283,2009.

[16] Peter A. Gloor,Jonas Krauss et.al, "Web Science 2.0: Identifying Trends through Semantic Social Network Analysis", International Conference on Computational Science and Engineering ,IEEE,PP.215-222,2009.

[17] Barahate Sachin R.,Shelake Vijay M, "A Survey and Future Vision of Data mining in Educational Field", Second International Conference on Advanced Computing & Communication Technologies,IEEE,PP.96-100,2012.

[18] David Combe,et.al, "Combining relations and text in scientific network clustering", IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, PP.1248-1253,2012.